Biostat 200C Homework 4

Due 11:59PM June 4th

Zian ZHUANG

Q1. Balanced one-way ANOVA random effects model

Consider the balanced one-way ANOVA random effects model with a levels and n observations in each level

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \ldots, a, \quad j = 1, \ldots, n.$$

where α_i are iid from $N(0, \sigma_\alpha^2)$, ϵ_{ij} are iid from $N(0, \sigma_\epsilon^2)$.

1. Derive the ANOVA estimate for μ, σ_α^2, and σ_ϵ^2. Specifically show that

$$\begin{aligned}
\mathbb{E}(\bar{y}_{\cdot\cdot}) &= \mathbb{E}\left(\frac{\sum_{ij} y_{ij}}{na}\right) = \mu \\
\mathbb{E}(\text{SSE}) &= \mathbb{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2\right] = a(n-1)\sigma_\epsilon^2 \\
\mathbb{E}(\text{SSA}) &= \mathbb{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2\right] = (a-1)(n\sigma_\alpha^2 + \sigma_\epsilon^2),
\end{aligned}$$

   which can be solved to obtain ANOVA estimate

$$\begin{aligned}
\widehat{\mu} &= \frac{\sum_{ij} y_{ij}}{na}, \\
\widehat{\sigma}_\epsilon^2 &= \frac{\text{SSE}}{a(n-1)}, \\
\widehat{\sigma}_\alpha^2 &= \frac{\text{SSA}/(a-1) - \widehat{\sigma}_\epsilon^2}{n}.
\end{aligned}$$

Answer:

(1).

$$
\begin{aligned}
\mathbb{E}(\bar{y}_{..}) &= \mathbb{E}\left(\frac{\sum_{ij} y_{ij}}{na}\right) \\
&= \frac{1}{na}\mathbb{E}\left(\sum_{ij} y_{ij}\right) \\
&= \frac{1}{na}\mathbb{E}\left(\sum_{ij} \mu + \alpha_i + \epsilon_{ij}\right) \\
&= \frac{1}{na}(\mathbb{E}\left(\sum_{ij} \mu\right) + \mathbb{E}\left(\sum_{ij} \alpha_i\right) + \mathbb{E}\left(\sum_{ij} \epsilon_{ij}\right)) \\
&= \frac{1}{na}(\mathbb{E}\left(\sum_{ij} \mu\right) + 0 + 0) \\
&= \frac{1}{na} na\mu \\
&= \mu
\end{aligned}
$$

(2).

$$
\begin{aligned}
\mathbb{E}(\text{SSE}) &= \mathbb{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}((\mu + \alpha_i + \epsilon_{ij}) - (\mu + \alpha_i + \bar{\epsilon}_{i.}))^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(\epsilon_{ij} - \bar{\epsilon}_{i.})^2\right] \\
&= \sum_{i=1}^{a}(n-1)\sigma_\epsilon^2, \quad \text{according to } E(x^2) = var(x) + E(x)^2 \\
&= a(n-1)\sigma_\epsilon^2
\end{aligned}
$$

(3).

$$
\begin{aligned}
\mathbb{E}(\text{SSA}) &= \mathbb{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}((\mu + \alpha_i + \bar{\epsilon}_{i\cdot}) - (\mu + \bar{\alpha}_{\cdot} + \bar{\epsilon}_{\cdot\cdot}))^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}((\mu + \alpha_i + \bar{\epsilon}_{i\cdot}) - (\mu + \bar{\alpha}_{\cdot} + \bar{\epsilon}_{\cdot\cdot}))^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(\alpha_i + \bar{\epsilon}_{i\cdot} - \bar{\alpha}_{\cdot} - \bar{\epsilon}_{\cdot\cdot})^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}((\alpha_i - \bar{\alpha}_{\cdot}) + (\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot}))^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(\alpha_i - \bar{\alpha}_{\cdot})^2 + (\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot})^2 + (\alpha_i - \bar{\alpha}_{\cdot})(\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot})\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(\alpha_i - \bar{\alpha}_{\cdot})^2 + (\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot})^2 + 0\right], \quad \text{given that } \alpha, \epsilon \text{ are independent} \\
&= n\mathbb{E}\left[\sum_{i=1}^{a}(\alpha_i - \bar{\alpha}_{\cdot})^2 + (\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot})^2\right] \\
&= n((a-1)\sigma_\alpha^2 + (a-1)\frac{\sigma_\epsilon^2}{n}), \quad \text{according to clt and } E(x^2) = var(x) + E(x)^2 \\
&= (a-1)(n\sigma_\alpha^2 + \sigma_\epsilon^2)
\end{aligned}
$$

2. Calculate the three estimates for the `pulp` example in class, check if your results match with the R output.

Answer:

$$
\begin{aligned}
\widehat{\mu} &= \frac{\sum_{ij} y_{ij}}{na} \\
&= \frac{1208}{4 * 5} \\
&= 60.4
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\sigma}_\epsilon^2 &= \frac{\text{SSE}}{a(n-1)} \\
&= \frac{1.70}{4(5-1)} \\
&= 0.10625
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\sigma}_\alpha^2 &= \frac{\text{SSA}/(a-1) - \widehat{\sigma}_\epsilon^2}{n} \\
&= \frac{1.34/(4-1) - 0.10625}{5} \\
&= 0.06808
\end{aligned}
$$

```r
#check with r output
data(pulp)
mean(pulp$bright)
```

```
## [1] 60.4
```

```r
(aovmod <- aov(bright ~ operator, data = pulp) %>% summary())
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## operator     3   1.34  0.4467   4.204 0.0226 *
## Residuals   16   1.70  0.1062
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#sigma_alpha
(aovmod[1][[1]][[3]][1] - aovmod[1][[1]][[3]][2]) / 5
```

```
## [1] 0.06808333
```

```r
#sigma_epsilon
aovmod[1][[1]][[3]][2]
```

```
## [1] 0.10625
```

Q2. Estimation of random effects

1. Assume the conditional distribution

$$\mathbf{y} \mid \boldsymbol{\gamma} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \sigma^2 \mathbf{I}_n)$$

and the prior distribution

$$\boldsymbol{\gamma} \sim N(\mathbf{0}_q, \boldsymbol{\Sigma}).$$

Then by the Bayes theorem, the posterior distribution is

$$f(\boldsymbol{\gamma} \mid \mathbf{y}) \quad = \quad \frac{f(\mathbf{y} \mid \boldsymbol{\gamma}) \times f(\boldsymbol{\gamma})}{f(\mathbf{y})},$$

where f denotes corresponding density. Show that the posterior distribution is a multivariate normal with mean

$$\mathbb{E}(\boldsymbol{\gamma} \mid \mathbf{y}) = \boldsymbol{\Sigma}\mathbf{Z}^T(\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Answer:

Given that,

$$y|\gamma \sim N(X\beta + Z\gamma, \sigma^2 I_n)$$
$$\gamma \sim N(0_q, \sum)$$

Then according to Bayes theorem, we can derive that,

$$
\begin{aligned}
f(\gamma|y) &\propto f(y,\gamma) = f(y|\gamma)f(\gamma) \\
&\propto \exp(-\frac{1}{2\sigma^2}(y-x\beta-z\gamma)^T(y-x\beta-z\gamma))\exp(-\frac{1}{2}\gamma^T(\sum)^{-1}\gamma) \\
&\propto \exp(-\frac{1}{2}(\frac{1}{\sigma^2}(\gamma^T z^T z\gamma - 2\gamma^T z^T y + 2\gamma^T z^T x\beta) + \gamma^T(\sum)^{-1}\gamma) \\
&\propto \exp(-\frac{1}{2}(\gamma^T(\frac{1}{\sigma^2}z^T z + (\sum)^{-1})\gamma - 2\gamma^T z^T(y-x\beta)))
\end{aligned}
$$

Then we found that this is the normal density (kernal) function. since we know that it follows that as a density must integrate to unity, then we know it follows normal density $\gamma|y \sim N(Mean_\gamma, Var_\gamma)$,

$$
\begin{aligned}
Var_\gamma &= (\frac{1}{\sigma^2}z^T z + (\sum)^{-1})^{-1} \\
Mean_\gamma &= (\frac{1}{\sigma^2}z^T z + (\sum)^{-1})^{-1} z^T(y-x\beta)\frac{1}{\sigma^2}
\end{aligned}
$$

Since we have,

$$
(E+FH^{-1}G)^{-1}FH^{-1} = E^{-1}F(H+GE^{-1}F)^{-1}
$$

Then we can plug in when $E = (\sum)^{-1}, F = z^T, G = z, H^{-1} = \frac{1}{\sigma}I$ and finally get,

$$
Mean_\gamma = \sum z^T(\frac{1}{\sigma^2}z^T z + (\sum)^{-1})^{-1}(y-x\beta)
$$

2. (**Optional**) For the balanced one-way ANOVA random effects model, show that the posterior mean of random effects is always a constant (less than 1) multiplying the corresponding fixed effects estimate.

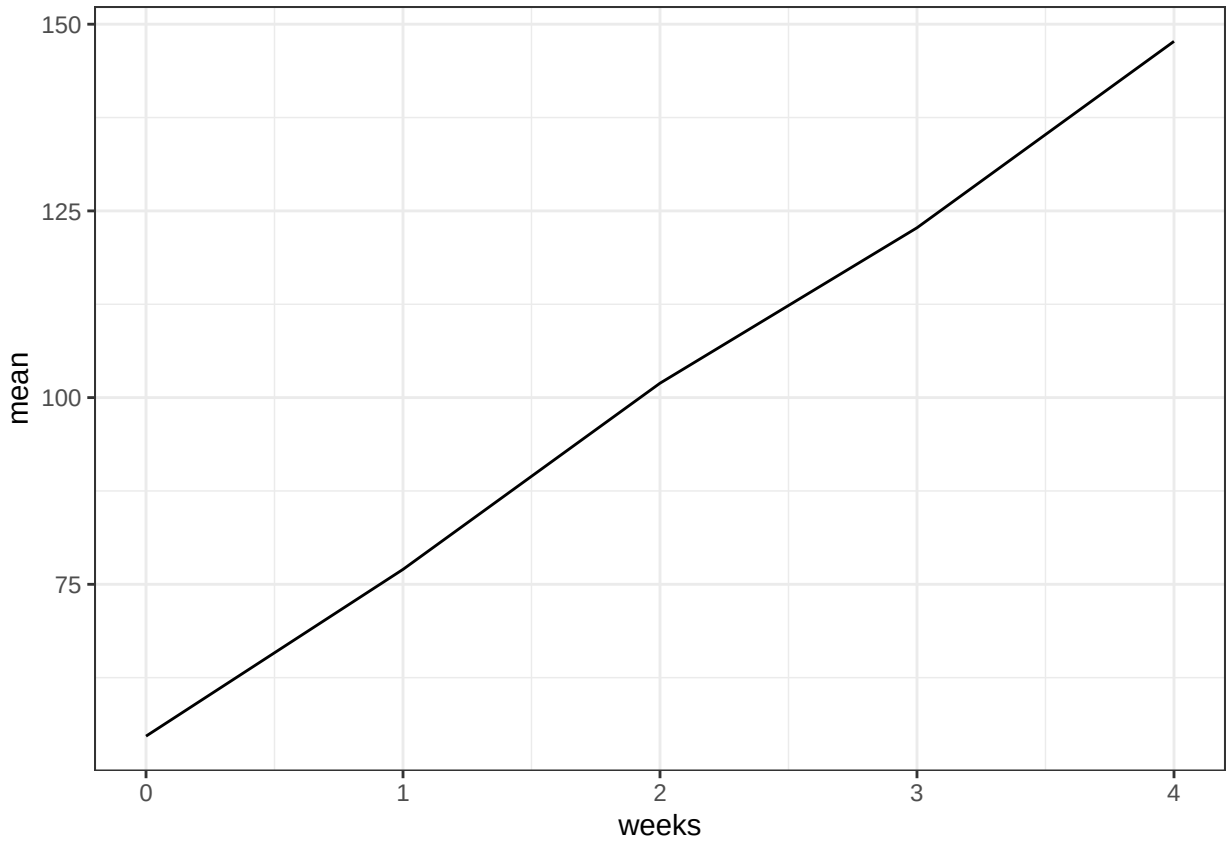
Q3. ELMR Exercise 11.1 (p251)

The `ratdrink` data consist of 5 weekly measurements of body weight for 27 rats. The first 10 rats are on a control treatment while 7 rats have thyroxine added to their drinking water and 10 rats have thiouracil added to their water.

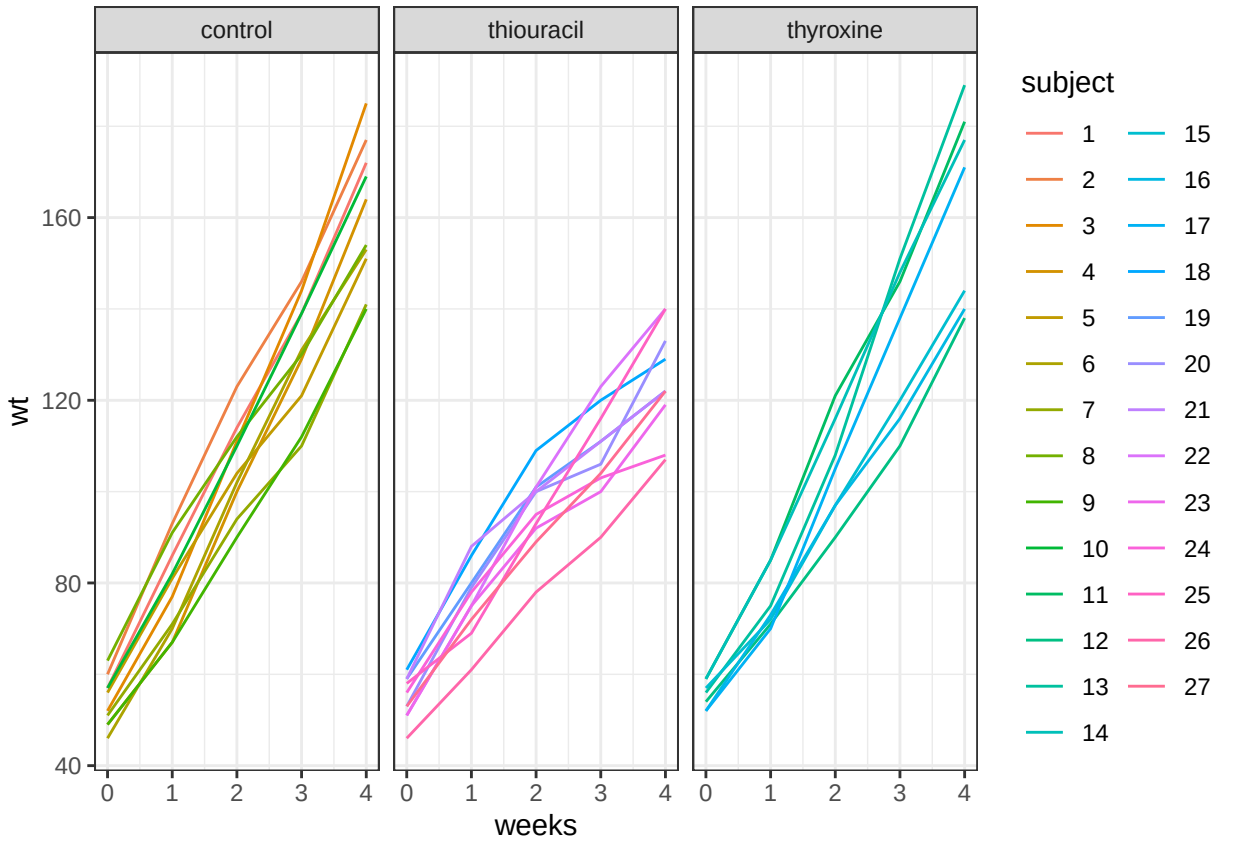
```
#help("ratdrink")
data(ratdrink)
```

1. Plot the data showing how weight increases with age on a single panel, taking care to distinguish the three treatment groups. Now create a three-panel plot, one for each group. Discuss what can be seen.

```
# general means
ratdrink %>%
  group_by(weeks) %>%
  summarise(mean=mean(wt)) %>%
  ggplot(.) + geom_line(aes(x=weeks,y=mean)) + theme_bw()
```




```
# details, grouped by treatments
ggplot(ratdrink) +
  geom_line(aes(weeks, wt, group=subject, color=subject)) +
  facet_wrap(~treat, ncol = 3) +
  theme_bw()
```



2. Fit a linear longitudinal model with a random slope and intercept for each rat. Each treatment group should have a different mean line. Give interpretation for the following estimates:

- The fixed effect intercept term.
- The interaction between thiouracil and week.
- The intercept random effect SD (standard deviation).

```
smod <- lmer(wt ~ treat*weeks + (1 + weeks | subject),
             data = ratdrink, REML = TRUE)
summary(smod)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: wt ~ treat * weeks + (1 + weeks | subject)
##    Data: ratdrink
##
## REML criterion at convergence: 878.7
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.83136 -0.54991  0.04003  0.58230  2.03660
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
```

```
##  subject  (Intercept) 32.49    5.700
##           weeks       14.14    3.760    -0.13
##  Residual             18.90    4.348
## Number of obs: 135, groups:  subject, 27
##
## Fixed effects:
##                       Estimate Std. Error t value
## (Intercept)            52.8800     2.0937  25.256
## treatthiouracil         4.7800     2.9610   1.614
## treatthyroxine         -0.7943     3.2628  -0.243
## weeks                  26.4800     1.2661  20.915
## treatthiouracil:weeks  -9.3700     1.7905  -5.233
## treatthyroxine:weeks    0.6629     1.9730   0.336
##
## Correlation of Fixed Effects:
##             (Intr) trtthr trtthy weeks  trtthr:
## treatthircl -0.707
## treatthyrxn -0.642  0.454
## weeks       -0.250  0.177  0.160
## trtthrcl:wk  0.177 -0.250 -0.113 -0.707
## trtthyrxn:w  0.160 -0.113 -0.250 -0.642  0.454
```

As we can tell from the results that:

- the average weight of a rat at week 0 is 52.88 among all treatment groups.
- for a rat in the thiouracil group, there will be an additional decrease in the average weight of -9.37 in a week.
- the average weight at week 0 has a standard deviation of 5.70 across all treatment groups

3. Check whether there is a significant treatment effect.

```
smod2 <- lmer(wt ~ weeks + (1 + weeks | subject),
              data = ratdrink, REML = TRUE)
KRmodcomp(smod, smod2)
```

```
## large : wt ~ treat * weeks + (1 + weeks | subject)
## small : wt ~ weeks + (1 + weeks | subject)
##          stat     ndf     ddf F.scaling   p.value
## Ftest  8.7125  4.0000 26.8141   0.94552 0.0001215 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

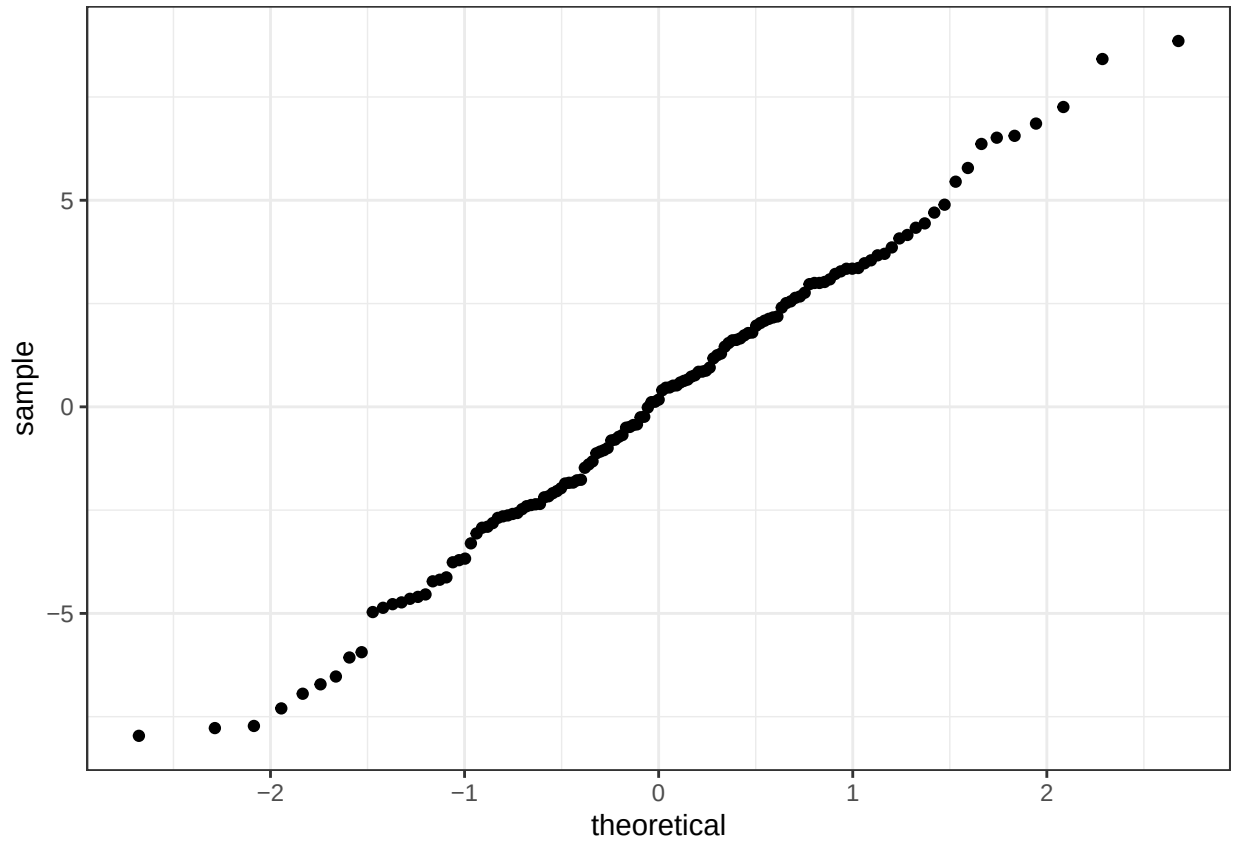
An approximate F-test based on the Kenward-Roger approach shows that the treatment has a significant effect ($p<.001$).

4. Construct diagnostic plots showing the residuals against the fitted values and a QQ plot of the residuals. Comment on the plots.

```
(diagd <- augment(smod))
```

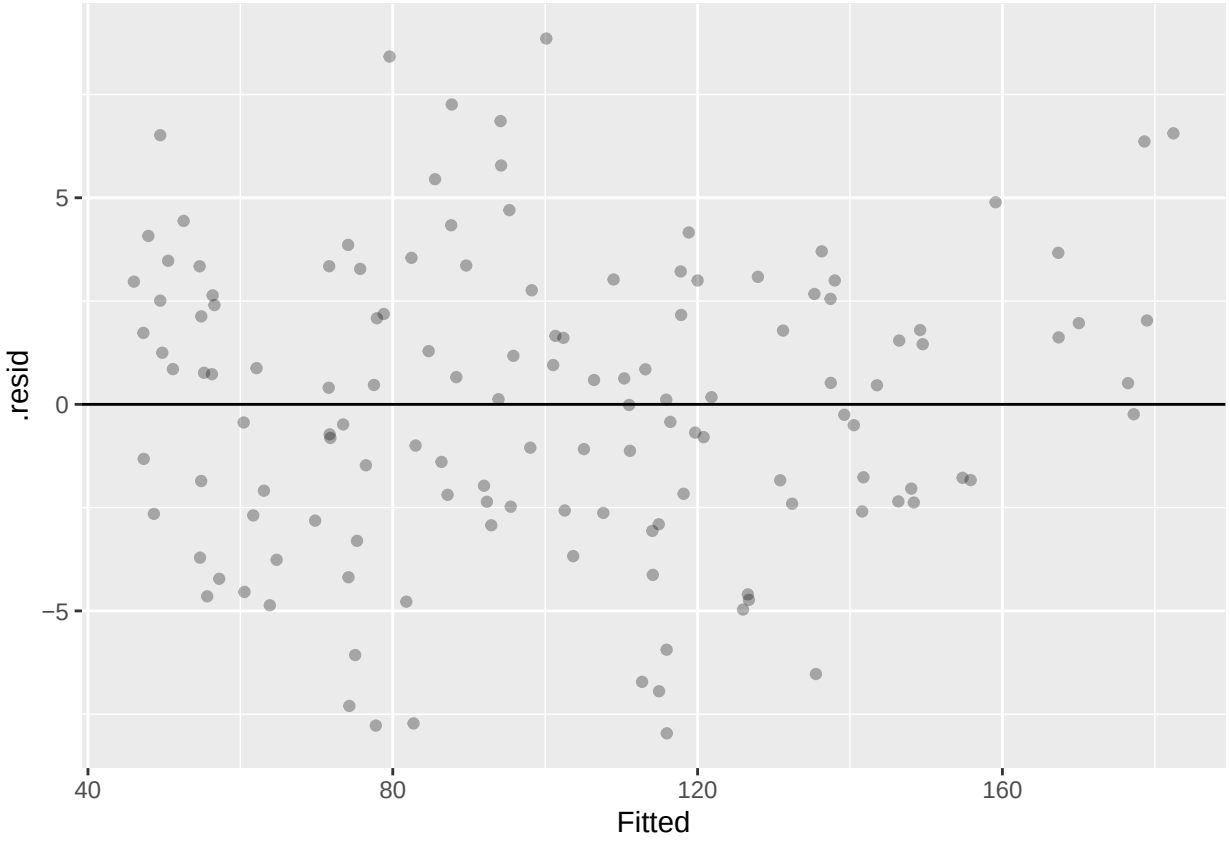
```
## # A tibble: 135 x 15
##       wt treat weeks subject .fitted .resid  .hat .cooksd .fixed   .mu .offset
##    <dbl> <fct> <int> <fct>     <dbl>  <dbl> <dbl>   <dbl>  <dbl> <dbl>   <dbl>
##  1    57 cont~     0 1          56.3  0.730 0.447 0.00686   52.9  56.3       0
##  2    86 cont~     1 1          84.7  1.29  0.236 0.00593   79.4  84.7       0
##  3   114 cont~     2 1         113.   0.848 0.183 0.00175  106.  113.        0
##  4   139 cont~     3 1         142.  -2.59  0.289 0.0338   132.  142.        0
##  5   172 cont~     4 1         170.   1.97  0.552 0.0937   159.  170.        0
##  6    60 cont~     0 2          60.4 -0.440 0.447 0.00249   52.9  60.4       0
##  7    93 cont~     1 2          89.6  3.36  0.236 0.0403    79.4  89.6       0
##  8   123 cont~     2 2         119.   4.16  0.183 0.0420   106.  119.        0
##  9   146 cont~     3 2         148.  -2.04  0.289 0.0209   132.  148.        0
## 10   177 cont~     4 2         177.  -0.240 0.552 0.00139  159.  177.        0
## # ... with 125 more rows, and 4 more variables: .sqrtXwt <dbl>, .sqrtrwt <dbl>,
## #   .weights <dbl>, .wtres <dbl>
```

```
# QQ plot of the residuals
diagd %>%
  ggplot(mapping = aes(sample = .resid)) +
  stat_qq() + theme_bw()
```




```
# Residuals vs fitted value plots.
diagd %>%
  ggplot(mapping = aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.3) +
```

```
  geom_hline(yintercept = 0) +
  labs(x = "Fitted", ylab = "Residuals")
```



We can tell from the results that the residuals follows a normal distribution. Nevertheless, it may have some underlying non-linear trend that is not being revealed.

5. Construct confidence intervals for the parameters of the model. Which random effect terms may not be significant? Is the thyroxine group significantly different from the control group?

```
confint(smod, method = "boot")
```

```
## Computing bootstrap confidence intervals ...

##
## 1 warning(s): Model failed to converge with max|grad| = 0.00342385 (tol = 0.002, component 1)

##                            2.5 %     97.5 %
## .sig01                 3.4742785  7.9308745
## .sig02                -0.5746331  0.4377225
## .sig03                 2.6954509  4.9893169
## .sigma                 3.6382536  4.9626449
## (Intercept)           48.9278842 57.0107629
## treatthiouracil       -1.2801973  9.8339961
## treatthyroxine        -7.2947238  5.5088111
```

```
## weeks                  24.0308433 28.7859447
## treatthiouracil:weeks -12.8092190 -5.4760023
## treatthyroxine:weeks   -3.0398378  4.9350908
```

As we can tell from the results that there is not a significant variance between the random intercept and the slope (95% CI .sig02 covers 0). Thyroxine group is not significantly different from the control group (`thyroxine:weeks` and `thyroxine` cover 0).